תרגול 8 - שיערוך פילוג בשיטות פרמטריות וסיווג גנרטיבי

Slides PDF Code

תקציר התיאוריה

הבעיה בגישה הלא פרמטרית

1. **Curse of dimensionality**: השיטות הלא פרמטריות לומדות את הפילוג בכל איזור על פי הדגימות שנמצאות באותו איזור באופן בלתי תלוי באיזורים האחרים במרחב. עקב כך, שיטות אלו דורשות כמות דגימות כזו שתכסה בצורה מספיק טובה את כל האיזורים הסבירים של מרחב הדגימות האפשריות. הבעיה היא שהגדול האפקטיבי של מרחב הדגימות גדל בצורה מעריכית עם המימד של הדגימות (האורך של הוקטור $\mathbf{x}$), לכן, בעבור מקרים שבהם המימד של דגימות הוא גדול, כמות הדגימות לרוב לא יחסכו בצורה טובה את המרחב ולכן השיערוך יהיה מאד לא מדוייק.
2. המודלים המתקבלים בשיערוך לא פרמטרי הם לרוב לא פונקציות שנוח לעבוד איתן. לדוגמא, על מנת לחשב את הצפיפות בנקודה מסויימת בעזרת KDE יש לבצע סכימה על כל הנקודות שנמצאות ב train set.

הגישה הפרמטרית

שיטה זו עושה שימוש במודלים פרמטרים בדומה לאופן שבו הדבר נעשה בגישה הדיסקרימינטיבית. בשיטה זו אנו נגביל את החיפוש של הפילוג למשפחה פרמטרית מסויימת, ונחפש את הפרמטרים האופטימאליים של המודל הנבחר. לרוב הפונקציה שאותה ננסה למדל הינה פונקציית צפיפות הפילוג (הPDF). חשוב לשים לב שבניגוד לשימוש במודלים פרמטרים בגישה הדטרמיניסטית, שם לא הייתה שום מגבלה על המודל הפרמטרי, כאן המודל חייב לייצר פילוג חוקי בעבור כל בחירה של פרמטרים (במקרה של PDF זה אומר פונקציה חיובית שאינטרגל עליה נותן 1)

ישנן שתי דרכים להתייחס לפרמטרים של המודל. שתי דרכים אלו מגיעים משתי גישות הקיימות בתחום של תורת השיערוך וכל גישה מובילה לדרך מעט שונה לבחירה של הפרמטרים האופטימאליים. בשתי הגישות אנו נסמן את וקטור הפרמטרים של המודל ב $\boldsymbol{\theta}$.

הגישה הלא-בייסיאנית (המכונה גם: קלאסית או תדירותית (Frequintist))

בגישה זו אנו נתייחס לפרמטרים באופן דומה לאופן שבו התייחסנו אליהם כאשר עסקנו בשיטות הדיסקרימינטיביות. תחת גישה זו אין כל העדפה של ערך מסויים של הפרמטרים על פני ערך אחר. את המודל הפרמטרי להסתברות / צפיפות הסתברות של משתנה אקראי $\mathbf{x}$ נסמן ב:

$$p_{\mathbf{x}}(\boldsymbol{x};\boldsymbol{\theta})$$

משערך MLE (Maximum Likelihood Estimator)

הדרך הנפוצה ביותר לבחור את הערך של $\boldsymbol{\theta}$ תחת הגישה הלא בייסיאנית היא בעזרת MLE. בשיטה זו נחפש את הערך של $\boldsymbol{\theta}$ אשר מסביר בצורה הכי טובה את המדגם הנתון. נסמן ב $p_{\mathcal{D}}(\mathcal{D};\boldsymbol{\theta})$ את ההסתברות לקבלת מדגם כל שהוא $\mathcal{D}=\{\boldsymbol{x}^{(i)}\}$. גודל זה מכונה ה**סבירות** (**likelihood**) של המדגם כפונקציה של $\boldsymbol{\theta}$. על מנת להדגיש את העובדה שהמדגם הוא למעשה גודל ידוע ואילו הגודל הלא ידוע שאותו נרצה לבדוק הינו $\boldsymbol{\theta}$, מקובל לסמן את פונקציית ה likelihood באפן הבא:

$$\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) \triangleq p_{\mathcal{D}}(D;\boldsymbol{\theta})$$

משערך ה MLE של $\boldsymbol{\theta}$ הוא הערך אשר ממקסם את פונקציית ה likelihood (או ממזער את המינוס שלה):

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} \ -\mathcal{L}(\boldsymbol{\theta};\mathcal{D})$$

כאשר הדגימות במדגם הן i.i.d (בעלות פילוג זהה ובלתי תלויות, כפי שנניח תמיד שמתקיים בבעיות supervised learning) נוכל להסיק כי:

$$p_{\mathcal{D}}(\mathcal{D};\boldsymbol{\theta}) = \prod_i p_{\mathbf{x}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

ולכן:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \ -\mathcal{L}(\boldsymbol{\theta};\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} \ -\prod_i p_{\mathbf{x}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

במקרים רבים נוכל להחליף את המכפלה על כל הדגימות בסכום, על ידי מקסימום של הלוג של פונקציית ה likelihood (בזכות המונוטוניות העולה של פונקציית ה log מובטח לנו שנקבל את אותם פרמטריים אופטימאלים):

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\min_{\boldsymbol{\theta}} \ -\log \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} \ -\sum_i \log\left(p_{\mathbf{x}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\right)$$

הגישה הבייסיאנית

בגישה זו אנו מניחים כי וקטור הפרמטרים $\boldsymbol{\theta}$ הינו ריאליזציה של וקטור אקראי בעל פילוג כלשהוא $p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. פילוג זה מכונה ה**פילוג הפריורי** (**prior distribution**) או ה**א-פריורי** (**a priori distribution**), זאת אומרת הפילוג של $\boldsymbol{\theta}$ לפני שראינו את המדגם. תחת גישה זו, המודל שלנו יהיה הפילוג של $\mathbf{x}$ **בהינתן** $\boldsymbol{\theta}$:

$$p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$$

משערך MAP (Maximum A-posteriori Probability)

הדרך הנפוצה ביותר לבחור את הערך של $\boldsymbol{\theta}$ תחת הגישה הבייסיאנית היא בעזרת MAP. בשיטה זו נחפש את הערך הכי סביר של $\boldsymbol{\theta}$ בהינתן המדגם $p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})$. פילוג זה מכונה ה**פילוג הפוסטריורי** (**posterior distribution**) או **א-פוסטריורי** (**a posteriori distribution**) (או הפילוג בדיעבד), זאת אומרת, הפילוג אחרי שראינו את המדגם.

אם כן, משערך ה MAP הוא וקטור הפרמטרים אשר ממקסמים את ההסתברות ה א-פוסטריורית:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} \ -p_{\boldsymbol{\theta}|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})$$

על פי חוק בייס, נוכל לכתוב זאת כ:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -\frac{p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p_{\mathcal{D}}(\mathcal{D})} = \arg\min_{\boldsymbol{\theta}} \ -p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta})p_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

כאשר הדגימות במדגם **בהינתן** $\boldsymbol{\theta}$ הן i.i.d מתקיים כי:

$$p_{\mathcal{D}|\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta}) = \prod_i p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

ולכן:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\prod_i p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})$$

גם כאן נוכל להפוך את המכפלה לסכום על ידי מזעור מינוס הלוג של הפונקציה:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \ -\log\left(p_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) - \sum_i \log\left(p_{\mathbf{x}|\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{\theta})\right)$$

(Linear Discriminant Analysis (LDA

LDA הינו אלגוריתם לפתרון בעיות סיווג בגישה גנרטיבית פרמטרית (לא בייסיאנית).

המודל הפרמטרי:

1. את הפילוג של $p_{\mathrm{y}}(y)$ נשערך ישירות מתוך התווית (זה פילוג דיסקרטי).
2. את הפילוג של $p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}|y)$ נמדל כפילוג נורמאלי.
3. אנו נניח כי מטריצת ה covariance של הפילוג הנורמאלי אינה תלויה בערך של y.

את מטריצת הקווריאנס של הפילוגים הנורמאליים (אותה נרצה לשערך) נסמן ב Σ. בנוסף, בעבור כל מחלקה c של y (הערכים שאותם הוא יכול לקבל) נסמן:

- $\mathcal{I}_c = \{i : \ y^{(i)} = c\}$ - זאת אומרת, אוסף האינדקסים של הדגמים במדגם שמקיימים $y^{(i)} = c$.
- $|\mathcal{I}_c|$ - מספר האינדקסים ב $\mathcal{I}_c$
- μ_c - וקטורי התוחלת של הפילוג הנורמאלי $p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}|c)$.

שיערוך של הפרמטריים בעזרת משערך MLE נותן את הפתרון הבא:

$$\boldsymbol{\mu}_c = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \boldsymbol{x}^{(i)}$$

$$\Sigma = \frac{1}{N} \sum_i \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}} \right) \left(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}} \right)^T$$

הפרדה לינארית

בעבור המקרה של סיווג בינארי (סיווג לשתי מחלקות) ושימוש בzero-one loss מתקבל החזאי הבא:

$$h\left(x\right) = \begin{cases} 1 & \boldsymbol{a}^T \boldsymbol{x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

כאשר:

$$\boldsymbol{a} = \Sigma^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \right)$$

$$b = \tfrac{1}{2} \left(\boldsymbol{\mu}_0^T \Sigma^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 \right) + \log \left(\frac{p_{\mathrm{y}}\left(1\right)}{p_{\mathrm{y}}\left(0\right)} \right)$$

נשים לב כי תנאי ההחלטה שבין שני התחומים הינו לינארי, ומכאן מקבל האלגוריתמם את שמו.

תרגיל 8.1 - שיערוך MLE

נתון מדגם $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$ של דגימות בלתי תלויות של משתנה אקראי x. מצאו את משערך ה MLE של המודלים הבאים:

1) פילוג נורמלי: $\mathrm{x} \sim N\left(\mu, \sigma^2\right)$ עם פרמטרים μ וσ^2 לא ידועים.

2) פילוג אחיד: $\mathrm{x} \sim U\left[0, \theta\right]$, עם פרמטר θ לא יודע.

3) פילוג אקספוננציאלי (**לקריאה עצמית**): $\mathrm{x} \sim \exp\left(\theta\right)$. עם פרמטר θ לא ידוע.

פיתרון 8.1

1)

המודל של פונקציית ה PDF יהיה:

$$p(x;\boldsymbol{\theta})) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

נסמן את וקטור הפרמטרים: $\boldsymbol{\theta} = [\mu, \sigma^2]^T$

משערך ה MLE נתון על ידי:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{MLE}} &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N}\log\left(p(x^{(i)};\boldsymbol{\theta})\right) \\ &= \arg\min_{\boldsymbol{\theta}} \quad -\sum_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi\theta_2}}\exp\left(-\frac{1}{2\theta_2}\left(x^{(i)}-\theta_1\right)^2\right)\right) \\ &= \arg\min_{\boldsymbol{\theta}} \quad \frac{N}{2}\log\left(2\pi\theta_2\right) + \sum_{i=1}^{N}\frac{1}{2\theta_2}\left(x^{(i)}-\theta_1\right)^2\end{aligned}$$

נפתור על ידי גזירה והשוואה ל 0 (נסמן ב $f(\boldsymbol{\theta})$ את פונקציית המטרה אותה יש למזער):

$$\begin{aligned}&\begin{cases}\frac{\partial}{\partial\theta_1}f(\boldsymbol{\theta}) = 0 \\ \frac{\partial}{\partial\theta_2}f(\boldsymbol{\theta}) = 0\end{cases} \\ \Leftrightarrow &\begin{cases}\sum_{i=1}^{N}\frac{1}{\theta_2}\left(x^{(i)}-\theta_1\right) = 0 \\ \frac{N}{2\theta_2} - \sum_{i=1}^{N}\frac{1}{2\theta_2^2}\left(x^{(i)}-\theta_1\right)^2 = 0\end{cases} \\ \Leftrightarrow &\begin{cases}\theta_1 = \frac{1}{N}\sum_{i=1}^{N}x^{(i)} \\ \theta_2 = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)}-\theta_1\right)^2\end{cases}\end{aligned}$$

מכאן ש:

$$\hat{\mu}_{\text{MLE}} = \hat{\theta}_1 = \frac{1}{N}\sum_{i=1}^{N}x^{(i)}$$

$$\hat{\sigma^2}_{\text{MLE}} = \hat{\theta}_2 = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)}-\hat{\mu}_{\text{MLE}}\right)^2$$

2)

המודל של פונקציית ה PDF יהיה:

$$p(x;\theta)) = \begin{cases}\frac{1}{\theta} & \theta \geq x_i \geq 0 \\ 0 & \text{else}\end{cases}$$

ולכן:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}}\prod_{i=1}^{N}p(x^{(i)};\boldsymbol{\theta}) = \begin{cases}\frac{1}{\theta^N} & \theta \geq x^{(i)} \quad \forall i \\ 0 & \text{else}\end{cases}$$

התנאי $\theta \geq x^{(i)}$ לכל i שקול ל $\theta > \max_i\{x^{(i)}\}$. מצד אחד נרצה לקיים תנאי זה בכדי שה likelihood לא יתאפס, מצד שני נרצה ש θ יהיה כמה שיותר קטן בכדי למקסם את $1/\theta^N$. לכן נבחר את ה θ מינימאלי אשר מקיים את התנאי:

$$\hat{\theta}_{\text{MLE}} = \max_i\{x^{(i)}\}$$

זאת אומרת, אנו נשערך את θ להיות הערך המסקימאלי במדגם.

3)

המודל של פונקציית ה PDF יהיה:

$$p(x;\boldsymbol{\theta})) = \theta \exp(-\theta x)$$

משערך ה MLE נתון על ידי:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{N} \log\left(p(x^{(i)};\boldsymbol{\theta})\right) \\ &= \arg\min_{\boldsymbol{\theta}} \; -N\log(\theta) + \theta\sum_{i=1}^{N} x^{(i)}\end{aligned}$$

נפתור על ידי גזירה והשוואה ל 0 (נסמן ב $f(\theta)$ את פונקציית המטרה אותה יש למזער):

$$\begin{aligned}&\frac{\partial}{\partial\theta} f(\theta) = 0 \\ \Leftrightarrow &-\frac{N}{\theta} + \sum_{i=1}^{N} x^{(i)} = 0 \\ \Leftrightarrow &\theta = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}}\end{aligned}$$

מכאן ש:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}}$$

תרגיל 8.2 - MAP

ביום טוב, עומרי כספי קולע בהסתברות p מהקו. ביום רע, הוא קולע בהסתברות q מהקו. α מהימים הם ימים טובים עבור עומרי.

ביום מסויים זרק עומרי N זריקות וקלע m מתוכם. מאמנו של עומרי צריך לזהות האם מדובר ביום טוב או רע של השחקן (ולהשאיר אותו או להחליף אותו בהתאמה).

מהו חוק ההחלטה אשר ממקסם את סיכויי המאמן לצדוק?

הניחו כי בהינתן המידע של האם יום מסויים הוא טוב או לא, ההסברות לקלוע זריקות שונות הינה הסתברות בלתי תלויה.

פתרון 8.2

נגדיר את המשתנים האקראיים והפילוגים שלהם:

- $\mathrm{x}^{(i)}$ - משתנה אקראי בינארי של האם עומרי קלע או לא בזריקה ה i-החטיא, 1-קלע)
- y - משתנה אקראי בינארי של האם היום הינו יום טוב או לא. (0-יום לא טוב, 1-יום טוב).

על פי הנתונים בשאלה:

$$p_{\mathrm{x}|\mathrm{y}}(x|0) = \begin{cases} 1-q & x=0 \\ q & x=1 \end{cases}$$

$$p_{\mathrm{x}|\mathrm{y}}(x|1) = \begin{cases} 1-p & x=0 \\ p & x=1 \end{cases}$$

$$p_{\mathrm{y}}(y) = \begin{cases} 1-\alpha & y=0 \\ \alpha & y=1 \end{cases}$$

בכדי למקסם את הסיכוי לחזות האם היום הוא יום טוב בהינתן המדגם נרצה למצוא איזה ערך יותר סביר בהינתן המדגם (יום טוב או רע), במילים אחרות אנו רוצים את ה y הכי סביר בהינתן $\mathcal{D} = \{x^{(i)}\}$:

$$\hat{y} = \arg\max_{y} p_{\mathrm{y}|\mathcal{D}}(y|\mathcal{D})$$

זוהי למעשה בעיית MAP קלאסית, כאשר y משמש למעשה כפרמטר בפילוג של $\mathrm{x}|\mathrm{y}$. בכדי לשמור על אחידות עם הסימונים שהגדרנו קודם לבעיות שיערוך נסמן את y ב θ. עלינו לפתור אם כן את:

$$\hat{\theta} = \arg\max_{\theta} p_{\theta|\mathcal{D}}(\theta|\mathcal{D}) = \arg\max_{\theta} p_{\mathcal{D}|\theta}(\mathcal{D}|\theta)p_{\theta}(\theta) = \arg\max_{\theta} p_{\theta}(\theta)\prod_{i} p_{\mathrm{x}|\theta}(x^{i}|\theta)$$

מכיוון ש θ יכול לקבל רק שני ערכים נוכל לבדוק את שניהם ולקבוע מי מהם סביר יותר.

בעבור $\theta = 0$ נקבל:

$$p_{\theta}(0)\prod_{i} p_{\mathrm{x}|\theta}(x^{(i)}|0) = (1-\alpha)q^{m}\left(1-q\right)^{N-m}$$

בעבור $\theta = 1$ נקבל:

$$p_{\theta}(1)\prod_{i} p_{\mathrm{x}|\theta}(x^{(i)}|1) = \alpha p^{m}\left(1-p\right)^{N-m}$$

לכן החיזוי האופטימאלי יהיה:

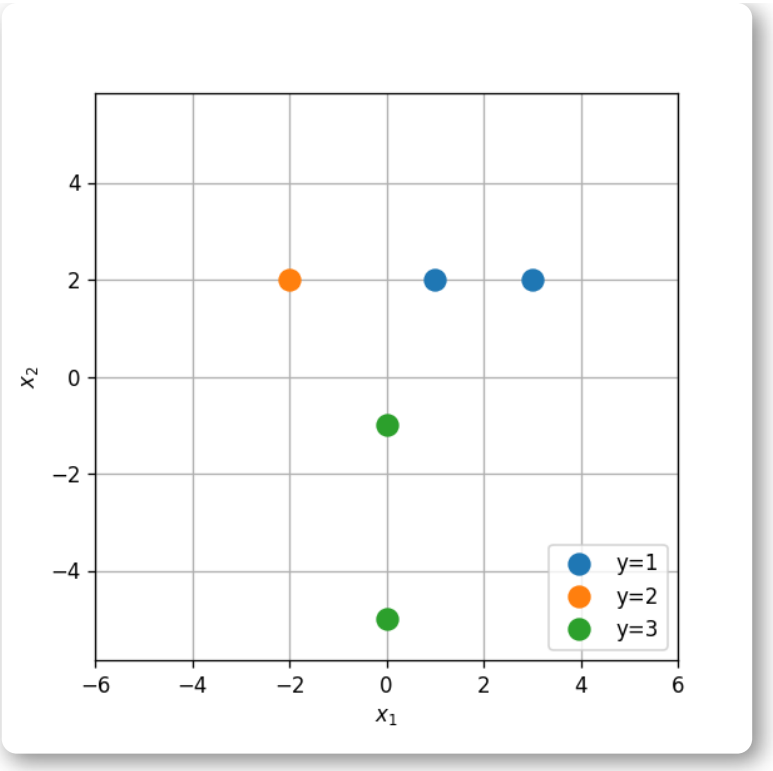
$$\begin{aligned}\hat{\theta} &= \begin{cases} 0 & (1-\alpha)q^{m}\left(1-q\right)^{N-m} > \alpha p^{m}\left(1-p\right)^{N-m} \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} 0 & \frac{1-\alpha}{\alpha}\left(\frac{q}{p}\right)^{m}\left(\frac{1-q}{1-p}\right)^{N-m} > 1 \\ 1 & \text{otherwise} \end{cases}\end{aligned}$$

תרגיל 8.3 - LDA

בסוואנה חיים שלושה זני פילים אשר נמצאים בסכנת הכחדה. ידוע כי כל אחד משלושת הזנים ניזון מצמחיה מעט שונה ועל מנת לשמר את אוכלוסיית הפילים מעוניינים לפזר להם אוכל ברחבי הסוואנה. בכדי למקסם את האפקטיביות של פעולה זו מעוניינים לשערך בכל נקודת חלוקה מהו הזן שהכי סביר להמצא באותה נקודה על מנת להתאים את סוג המזון לזן זה.

הפילוג של זני הפילים על פני הסוואנה אינו ידוע אך נתונות לנו התצפית הבאה של הקואורדינטות בהם נצפו הפילים:

Type	x_1	x_2
1	1	2
1	3	2
2	-2	2
3	0	-1
3	0	-5



השתמש במסווג LDA על מנת לבנות חזאי אשר ישערך את הזן הנפוץ ביותר בכל קואורדינטה.

פתרון 8.3

נחשב את הפרמטרים של המודל הפרמטרי של LDA.

נסמן ב $\mathcal{I}_c$ את אוסף כל התצפיות שבהם הזן הוא c:

$$\mathcal{I}_1 = \{1, 2\}$$

$$\mathcal{I}_2 = \{3\}$$

$$\mathcal{I}_3 = \{4, 5\}$$

נשערך את $p_{\mathrm{y}}(y)$:

$$p_{\mathrm{y}}(y) = \begin{cases} \frac{|\mathcal{I}_1|}{N} = \frac{2}{5} & 1 \\ \frac{|\mathcal{I}_2|}{N} = \frac{1}{5} & 2 \\ \frac{|\mathcal{I}_3|}{N} = \frac{2}{5} & 3 \end{cases}$$

נחשב את התוחלות של כל אחת משלושת הפילוגים $p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}|c)$:

$$\boldsymbol{\mu}_1 = \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \boldsymbol{x}^{(i)} = \frac{1}{2} \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right) = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \boldsymbol{x}^{(i)} = \begin{pmatrix} -2 \\ 2 \end{pmatrix}$$

$$\boldsymbol{\mu}_3 = \frac{1}{|\mathcal{I}_3|} \sum_{i \in \mathcal{I}_3} \boldsymbol{x}^{(i)} = \frac{1}{2} \left(\begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ -5 \end{pmatrix} \right) = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$$

נחשב את מטריצת covariance המשותפת של הפילוגים:

$$\Sigma = \frac{1}{N}\sum_i (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})^T$$

דרך נוחה לחשב את הסכום בביטוי זה הינה באופן הבא. נגדיר את המטריצה של התצפיות לאחר חיסור של התוחלת המתאימה לכל זן:

$$\tilde{X} = \begin{pmatrix} -\boldsymbol{x}_1- \\ -\boldsymbol{x}_2- \\ -\boldsymbol{x}_3- \\ -\boldsymbol{x}_4- \\ -\boldsymbol{x}_5- \end{pmatrix} - \begin{pmatrix} -\boldsymbol{\mu}_{y_1}- \\ -\boldsymbol{\mu}_{y_2}- \\ -\boldsymbol{\mu}_{y_3}- \\ -\boldsymbol{\mu}_{y_4}- \\ -\boldsymbol{\mu}_{y_5}- \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ -2 & 2 \\ 0 & -1 \\ 0 & -5 \end{pmatrix} - \begin{pmatrix} 2 & 2 \\ 2 & 2 \\ -2 & 2 \\ 0 & -3 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 2 \\ 0 & -2 \end{pmatrix}$$

ניתן להראות כי ניתן לכתוב את הסכום בביטוי ל Σ באופן הבא:

$$\begin{aligned} \Sigma &= \frac{1}{N}\sum_i (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})^T = \frac{1}{N}\tilde{X}^T\tilde{X} \\ &= \frac{1}{5}\begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 \end{pmatrix}\begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 2 \\ 0 & -2 \end{pmatrix} \\ &= \frac{1}{5}\begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix} \end{aligned}$$

נשתמש כעת בפילוגים שאותם שיערכנו על מנת לבנות את החזאי. האיזור שבו זן 1 הינו הזן הסביר ביותר הינו האיזור שבו מתקיים:

$$\begin{cases} p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x}) > p_{\mathrm{y}|\mathbf{x}}(2|\boldsymbol{x}) \\ p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x}) > p_{\mathrm{y}|\mathbf{x}}(3|\boldsymbol{x}) \end{cases}$$

נחשב את התנאי הראשון

$$\begin{aligned} & p_{\mathrm{y}|\mathbf{x}}(1|\boldsymbol{x}) > p_{\mathrm{y}|\mathbf{x}}(2|\boldsymbol{x}) \\ \Leftrightarrow\ & p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}|1)p_{\mathrm{y}}(1) > p_{\mathbf{x}|\mathrm{y}}(\boldsymbol{x}|2)p_{\mathrm{y}}(2) \\ \Leftrightarrow\ & \frac{1}{\sqrt{4\pi^2|\Sigma|}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}p_{\mathrm{y}}(1) > \frac{1}{\sqrt{4\pi^2|\Sigma|}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}p_{\mathrm{y}}(2) \\ \Leftrightarrow\ & -\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \log\left(p_{\mathrm{y}}(1)\right) > -\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \log\left(p_{\mathrm{y}}(2)\right) \\ \Leftrightarrow\ & \boldsymbol{x}^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \tfrac{1}{2}(\boldsymbol{\mu}_2^T\Sigma^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\Sigma^{-1}\boldsymbol{\mu}_1) + \log\left(\frac{p_{\mathrm{y}}(1)}{p_{\mathrm{y}}(2)}\right) > 0 \end{aligned}$$

זוהי למעשה הפרדה לשני תחומים על ידי הקו הבא:

$$\boldsymbol{a}^T\boldsymbol{x} + b = 0$$

כאשר:

$$\begin{aligned} \boldsymbol{a} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} 10 \\ 0 \end{pmatrix} \\ b &= \tfrac{1}{2}(\boldsymbol{\mu}_2^T\Sigma^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\Sigma^{-1}\boldsymbol{\mu}_1) + \log\left(\frac{p_{\mathrm{y}}(1)}{p_{\mathrm{y}}(2)}\right) = \log(2) \end{aligned}$$

זוהי כמובן התוצאה עבור מסווג LDA בינארי בין שני הזנים של $\mathrm{y} = 1$ ו $\mathrm{y} = 2$.

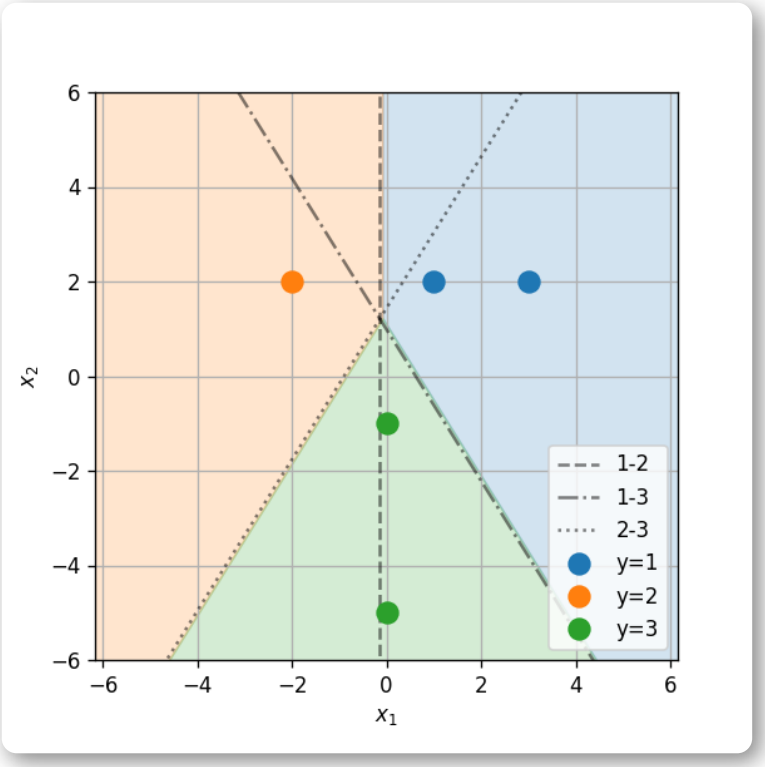
מכאן שקו המפריד בין זן 1 ל זן 2 נתון על ידי:

$$1-2: \quad 10x_1 + \log(2) = 0$$

באופן דומה ניתן לחשב גם את שני קווי ההפרדה האחרים (בין 1 ל 3 ובין 2 ל 3):

$$1-3: \quad 5x_1 + \frac{25}{8}x_2 + \frac{55}{16} = 0$$

$$2-3: \quad -5x_1 + \frac{25}{8}x_2 + \frac{55}{16} - \log(2) = 0$$



תרגיל מעשי - שיערוך הפילוג של זמני נסיעה בניו יורק

Code

נחזור לבעיה מהתרגול הקודם של שיערוך הפילוג של זמן הנסיעה של מונית מתוך מדגם הבא:

ay of ek	duration	dropoff northing	dropoff easting	pickup northing	pickup easting	tip amount	fare amount	payment type	trip distance	passenger count	
3	11.5167	4515.18	588.155	4512.98	586.997	0	9.5	2	2.76806	2	0
6	12.6667	4512.63	584.85	4512.92	587.152	0	10	2	3.21868	1	1
0	5.51667	4513.17	585.434	4513.36	587.005	2.49	7	1	2.57494	1	2
1	9.88333	4512.55	586.672	4511.73	586.649	1.65	7.5	1	0.965604	1	3
2	8.68333	4511.76	585.262	4511.89	586.967	1.66	7.5	1	2.46229	1	4
3	9.43333	4511.54	585.169	4512.88	585.926	2.2	7.5	1	1.56106	5	5
5	7.95	4514.21	588.71	4515.08	586.731	1	8	1	2.57494	1	6

ay of ek	duration	dropoff northing	dropoff easting	pickup northing	pickup easting	tip amount	fare amount	payment type	trip distance	passenger count	
5	4.95	4509.55	585.844	4509.71	585.345	0	5	2	0.80467	1	7
5	11.0667	4507.74	583.671	4509.48	585.422	1.1	10	1	3.6532	1	8
3	4.21667	4513.71	587.701	4514.93	587.875	1.36	5.5	1	1.62543	6	9

ננסה להתאים מודל פרמטרי בעזרת שיערוך MLE.

ניסיון 1: פילוג גאוסי

נשתמש במודל של פילוג נורמלי לתיאור הפילוג של משך הנסיעה. למודל זה שני פרמטרים, התוחלת μ והשונות σ .

סימונים והנחות:

- N - מספר הדגמים במדגם.
- $\boldsymbol{\theta} = [\mu, \sigma]^T$ - וקטור הפרמטרים של המודל
- $p_{\text{normal}}\left(x_i;\boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right), i = 1, ..., N$ - המודל

ראינו כי בעבור המודל הנורמלי, ניתן למצוא את הפרמטרים של משערך הMLE באופן מפורש (אנליטית), והפתרון נתון על ידי:

$$\mu = \frac{1}{N}\sum_i x_i$$

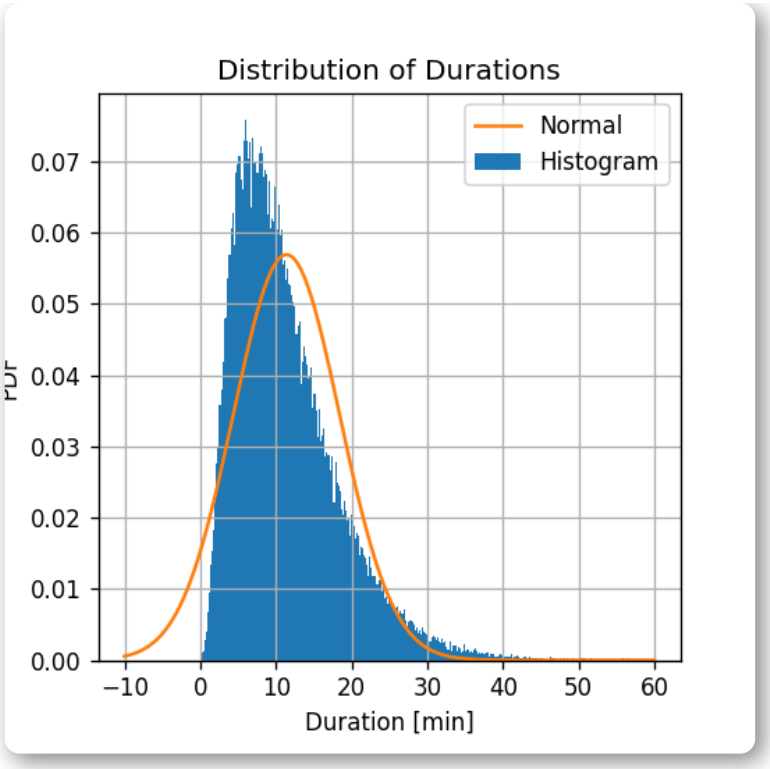
$$\sigma = \sqrt{\frac{1}{N}\sum_i \left(x_i - \mu\right)^2}$$

בעבור המדגם הנתון נקבל:

$$\hat{\mu} = 11.4 \text{ min}$$

$$\hat{\sigma} = 7.0 \text{ min}$$

נשרטט את ההיסטוגרמה של של משכי הנסיעה יחד עם הפילוג הנורמלי המשוערך:



נראה כי הפילוג הנורמלי נותן קירוב מאד גס לפילוג האמיתי. במקרים רבים קירוב זה יהיה מספיק, אך במקרה זה ננסה לשפר את השיערוך שלנו.

עובדה אחת שמאד מטרידה לגבי הפילוג שקיבלנו הינה שישנו סיכוי לא אפסי לקבל נסיעות עם משך נסיעה שלילי.

ננסה להציע מודל טוב יותר.

נסיון 2: פילוג Rayleigh

פילוג Rayleigh מתאר את הפילוג של האורך האוקלידי (l_2 norm) של וקטור גאוסי דו מימדי עם תוחלת 0 וחוסר קורלציה ופילוג זהה לשני רכיבי הוקטור. במלים אחרות, עבור וקטור בעל הפילוג הבא:

$$\boldsymbol{Z} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2 & 0\\0 & \sigma^2\end{bmatrix}\right)$$

פילוג Rayleigh מתאר את הפילוג של הגודל $\|\boldsymbol{Z}\|_2 = \sqrt{Z_x^2 + Z_y^2}$.

פונקציית צפיפות ההסתברות של פילוג Reyligh נתונה על ידי:

$$p_{\text{Rayleigh}}(z;\sigma) = \frac{z}{\sigma^2}\exp\left(-\frac{z^2}{2\sigma^2}\right), \quad z \geq 0$$

נשים לב כי הפילוג מוגדר רק בעבור ערכים חיוביים. לפילוג זה פרמטר יחיד σ שנקרא פרמטר סקאלה (scale parameter). בניגוד לפילוג הנורמלי, פה σ אינה שווה לסטיית התקן של הפילוג.

ניתן מוטיבציה קצרה לבחירה שלנו במודל זה.

מוטיבציה לשימוש בפילוג Rayleigh

נתחיל עם ההנחה שוקטור המחבר את נקודת תחילת הנסיעה עם נקודת סיום הנסיעה הינו וקטור דו מימדי אשר מפולג נרמלית ולשם הפשטות נניח כי רכיביו מפולגים עם פילוג זהה וחסר קורלציה.

בנוסף לשם הפשטות נניח כי המונית נוסעת בקירוב בקו ישר בין נקודת ההתחלה והסיום ולכן המרחק אותו נוסעת המכונית יהיה מפולג על פי פילוג Reyleigh. נניח בנוסף כי מהירות הנסיעה קבוע ולכן משך הנסיעה פורפורציוני למרחק ולכן גם הוא יהיה מפולג על פי פילוג Reyleigh.

חישוב

לשם השלמות נסמן את וקטור הפרמטרים של ב: $\theta = [\sigma]$

במקרה זה המודל נתון על ידי:

$$p_{\text{rayleigh}}\left(\boldsymbol{x};\theta\right) = \prod_{i=1}^{N} \frac{x_i}{\theta^2} \exp\left(-\frac{x_i^2}{2\theta^2}\right)$$

ופונקציית ה log likelihood תהיה:

$$\begin{aligned} l_{\text{rayleigh}}\left(\theta\right) &= \sum_i \log\left(p_{\text{rayleigh}}\left(x_i;\theta\right)\right) \\ &= \sum_i \log\left(x_i\right) - 2N\log\left(\theta\right) - \frac{1}{2\theta^2}\sum_i x_i^2 \end{aligned}$$

בעיית האופטימיזציה שלנו תהיה:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \quad -\sum_i \log\left(x_i\right) + 2N\log\left(\theta\right) + \frac{1}{2\theta^2}\sum_i x_i^2$$

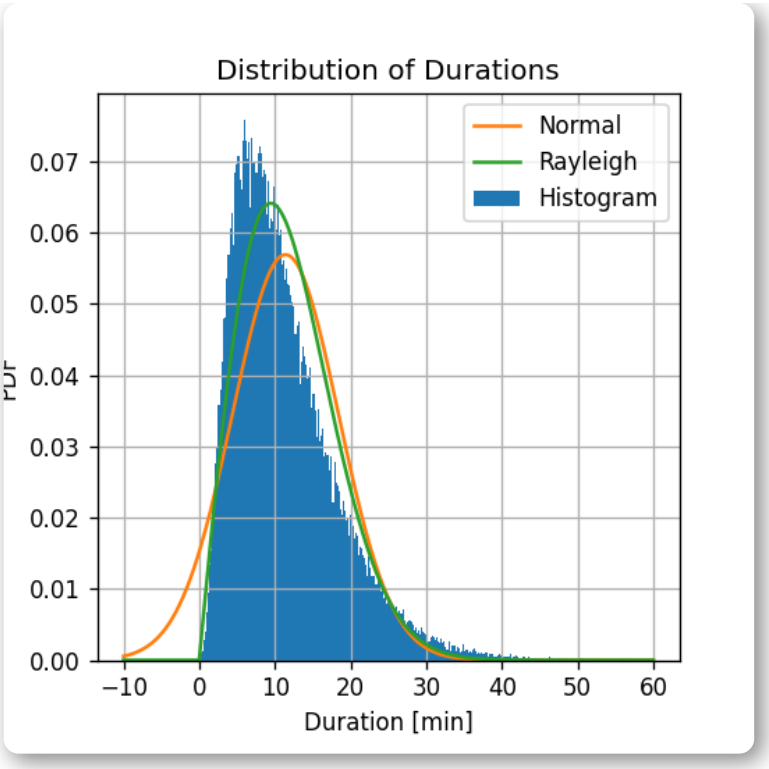
גם בעבור המקרה הזה נוכל לפתור את בעיית האופטימיזציה באופן אנליטי על ידי גזירה והשוואה לאפס:

$$\begin{aligned} &\frac{\partial l_{\text{rayleigh}}\left(\theta\right)}{\partial\theta} = 0 \\ \Leftrightarrow &-\frac{2N}{\theta} + \frac{\sum_i x^2}{\theta^3} = 0 \\ \Leftrightarrow &\hat{\sigma} = \theta = \sqrt{\frac{1}{2N}\sum_i x^2} \end{aligned}$$

בעבור המדגם הנתון נקבל:

$$\hat{\sigma} = 9.5$$

נוסיף את השיערוך החדש שקיבלנו לגרף ממקודם:



על פי הדמיון בין ההיסטוגרמה לפונקציות הפילוג ששיערכנו, נראה כי המודל של פילוג Rayleigh נותן תוצאה מעט יותר טובה מהמודל הנורמלי, בנוסף ניתן לראות גם כי כעת אין הסתברות שונה מ0 לקבל משך נסיעה שלילי.

ננסה מודל נוסף.

נסיון 3: Generalized Gamma Distribution

פילוג Rayleigh הינו מקרה פרטי של משפחה כללית יותר של פונקציות פילוג המכונה Generalized Gamma Distribution. פונקציית צפיפות ההסתברות של משפחה זו נתונה על ידי:

$$p_{\text{gengamma}}\left(z;\sigma,a,c\right)=\frac{cz^{ca-1}\exp\left(-\left(z/\sigma\right)^{c}\right)}{\sigma^{ca-1}\Gamma\left(a\right)},\quad z\geq 0$$

(כשאר Γ היא פונקציה המוכנה פונקציית גמא (gamma function))

למודל זה 3 פרמטרים: $\boldsymbol{\theta}=\left[\sigma,a,c\right]^{T}$.

בעבור $c=2$ ו $a=1$ נקבל את פילוג Rayleight כאשר $\sigma_{gamma}=2\sigma_{rayleigh}$.

בניגוד למקרים של פילוג נורמלי ופילוג Rayleigh, במקרה זה לא נוכל למצוא בקלות את הפרמטרים האופטימאלים של המשערך באופן אנליטי. לכן, לשם מציאת הפרמטרים נאלץ להעזר בפתרון נומרי. בפועל נעשה שימוש באחת החבילה של Python הנקראת SciPy. חבילה זו מכילה מודלים הסברותיים רבים ומכילה מספר רב של כלים הקשורים למודלים אלו, כגון מציאת הפרמטרים האופטימאלים בשיטת MLE על סמך מדגם נתון. את הפונקציות הקשורות למודל הGeneralized Gamma Distribution ניתן למצוא כאן.

אתם תעשו שימוש בפונקציות אלו בתרגיל הבית הרטוב.

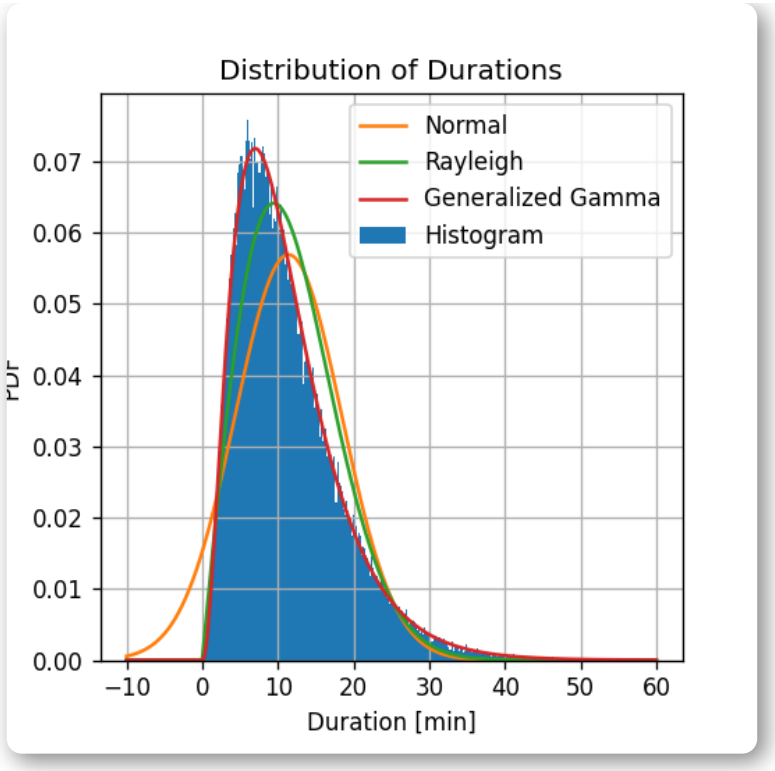
שימוש בפונקציה הנ"ל, מניב את התוצאות הבאות:

$$\hat{a}=4.4$$

$$\hat{c}=0.8$$

$$\hat{\sigma}=1.6$$

נוסיף את השיערוך החדש שקיבלנו לגרף הקודם:



ניתן לראות המודל של Generalized Gamma Distribution אכן מניב תוצאה אשר דומה מאד לצורת ההסטוגרמה.